10/652,070
MV-892

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 590 858 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **93307488.2**

(51) Int. Cl.[5] : **G06F 15/403**

(22) Date of filing : **22.09.93**

(30) Priority : **29.09.92 US 953166**

(43) Date of publication of application : **06.04.94 Bulletin 94/14**

(84) Designated Contracting States : **DE FR GB**

(71) Applicant : **XEROX CORPORATION**
Xerox Square
Rochester New York 14644 (US)

(72) Inventor : **Henderson, Richard D.**
505 Aleta Avenue
San Jose, California 95128 (US)
Inventor : **Barbarino, Michael J.**
363 California Street
Moss Beach, California 94038 (US)

(74) Representative : **Goode, Ian Roy et al**
Rank Xerox Patent Department, Albion House,
55-59 New Oxford Street
London WC1A 1BS (GB)

(54) Method for performing a search of a plurality of documents for similarity to a query.

(57) A method for performing a search of a plurality of documents for similarity to a query word includes retrieving a first document (20), and determining (21,23) a number of occurrences of the at least one query word in the first document. Then, a next document is retrieved (25) and a number of occurrences of the at least one query word in the next document is determined (27,28). The steps are repeated (30) until each of the plurality of documents have been retrieved, and the number of occurrences of the at least one query word has been determined in each of the plurality of documents. The query word can include a plurality of query words, all of which are searched in each document, in turn, rather than being searched word by word in the whole collection of documents. The documents are then ranked according to the number of occurrences of the query words determined in each document, and a list of documents is produced according to the document ranking.

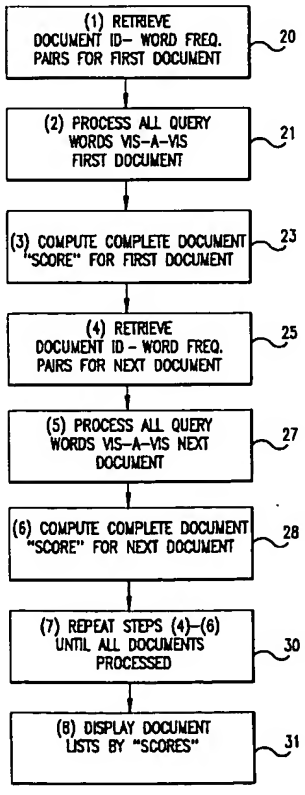

FIG.2

EP 0 590 858 A1

This invention relates to improvements in text and image processing methods and techniques, and more particularly to improvements in methods for word or term identification and location in document images, and still more particularly to improvements in methods for computer searching a number of document images for existence of query words or terms with reduced memory requirements.

There has been increasingly widespread interest in document processing, both in electronic and in paper document forms. Often it is desired to locate particular search terms within a large corpus of documents; for example, in performing research to locate papers or publications that pertain to particular subjects, in finding particular testimony in deposition or discovery documents that contain particular words or phrases, in locating relevant court decisions in a legal database that have certain key words, and in manifold other instances.

Sometimes the documents are presented in electronic form in which the document text and images have been encoded in an electronic memory media from which the documents can be retrieved for perusal or for "hard copy" or paper reproduction. In the past, when a large number of such documents are to be searched to locate one or more query terms, usually words, an index is built against which the query terms are compared. Such index generally is formed of two parts. The first part is a document identifier (herein the "document id"). The document id is merely an identification of each document in the collection, and may be a number, key word or phrase, or other unique identifier. The second part is a word and the number of times the word appears in the document with which it is identified (herein the "word frequency").

In the past, as shown in Figure 1, to identify the particular documents in which search or query words exist, usually the index of all of the words is brought into a computer memory 10, and the query words are compared, one at time, against each of the words in the memory. As each word is compared, a "score" is kept of the documents in which it appears. Thus, a first query word is processed 11, and a partial "score" is computed 13 for the first word. Then a next query word is processed 14, and a cumulative "score" is computed 16. As the successive query words and cumulative scores are processed until completed 17, the cumulative score is continued to be generated. After the last query word has been searched, the "scores" can be used to identify or sort the documents 18 in order of the number of "hits" by the query words, and a list of documents found can be displayed 19.

Such techniques, however, require a large amount of computer accessible memory, particularly for large document collections. The memory requirement often makes it impractical for document searching on personal or portable computers, even if the documents are stored on large capacity memory disks, and generally require large, mainframe computers with associated large memories.

In the field of image processing, recently, direct paper document searching techniques have been proposed in which one or more morphological properties of the images on the document are processed and used for comparison against a query word, term or image. In accordance with such techniques, a document is scanned and the morphological properties of its various images directly determined without decoding the content of the image. In performing searches of a large corpus of documents, however, one technique that can be used is to generate an index similar to that described above, but with a list of frequencies of morphological properties used in place of the words. Again, especially in large document collections, a large amount of memory is required to perform search queries.

In light of the above, it is, therefore, an object of the invention to provide an improved method for performing a similarity search on a large collection of documents using less memory than conventional methods heretofore employed.

It is another object of the invention to provide an improved method of the type described that can be performed efficiently.

The present invention provides a method of performing a search of a plurality of documents, according to claims 1, 2, 4, 5 and 6 of the appended claims.

In accordance with a broad aspect of the invention, a method for performing a search of a plurality of documents for similarity to a query term or word is presented. The method includes retrieving a first document, and determining a number of occurrences of the query word in the first document. The method then includes retrieving a next document and determining a number of occurrences of the query word in the next document. The steps are repeated until each of the plurality of documents have been retrieved, and the number of occurrences of the query word has been determined in each of the plurality of documents.

The query word can include a plurality of query terms, all of which are searched in each document, in turn, rather than being searched term by term in the whole collection of documents. The documents are then ranked according to the number of occurrences of the query words determined in each document, and a list of documents is produced according to the document ranking.

In one embodiment, a list of words contained within the retrieved document is generated, and the query words are compared to the generated list of words.

In another embodiment, all of the query words are compared against a first portion of the documents. Subsequently, all of the query words are compared against a second portion of the documents.

The documents are then ranked, according to the number of occurrences of the query words determined in each document, and a list of the documents is generated according to the document ranking.

In another embodiment, the documents are organized into an inverted index. In this case, instead of retrieving a document, the segment of a list of document-id and term-frequency pairs related to the query term and the document is examined.

The present invention further provides a programmable document searching system when suitably programmed for carrying out the method of any of claims 1 to 10.

The invention is illustrated in the accompanying drawing, in which:

Figure 1 is a block diagram outlining the steps for performing a similarity search of a corpus of documents, in accordance with the prior art; and

Figure 2 is a block diagram outlining the steps for performing a similarity search of a corpus of documents in accordance with a preferred embodiment of the invention.

The present invention relates to a method for performing a search of a plurality of documents, which method may be carried out in any conventional information processing system, such as that schematically illustrated in, and described with reference to, Fig. 1 of European patent application 93306281.2, a copy of which was filed with the present application.

This invention relates to techniques for performing similarity searches of the type in which the similarity search is performed with a query formed of a sequence of one or more words, syllables, phrases, images, or the like. Although the term "query word" is used herein, it should be understood that the "word" refers to a word, a word portion, or portions of a document or image which comprises letters, numbers, or other language symbols including non-alphabetic linguistic characters such as ideograms or foreign syllabries, and word or character substitutes, such as "wildcard" characters or the like. The result of the similarity search is a ranked list of documents from the indexed collection that have the highest similarity quotient to the query. The similarity quotient of a document with regard to a query is a number that results from a user defined formula that may include the number of documents in which each query word appears, the number of times it appears in each document, and the number of documents in the corpus. In some instances, it may be desirable to include different weights to be applied that designate relative importance of query words, or order of appearance of query words, or other similar search criterion.

In order to accomplish the similarity search according to the invention, an inverted index is preferably used. The inverted index contains a list of pairs of document identifiers and word frequency for each unique word in the corpus, or collection of documents. The word frequency is the number of times the word appears in the document identified by the document id with which it is paired. The document id - word frequency pairs are preferably arranged in ascending or descending order by document id.

The method of the invention is in contrast to previous methods in which the calculation of a similarity quotient is usually made by going entirely through the list of pairs of document identifiers and word frequencies for a single query word, and as each query word is being processed, computing partial scores for each document found in the list. In accordance with the method of a preferred embodiment of the invention, with reference now to Figure 2, rather than accessing all the document id - word frequency pairs for a query word before accessing those of another query word, the comparison is switched from one stream of document - word frequency pairs to another. Thus, all the document id - word frequency pairs for one document are visited before going on to others.

Accordingly, the document id - word frequency pairs for the first document are retrieved 20 into a computer memory. Thus, it will be appreciated that the technique of the invention is particularly well suited for use in memory constrained cases, and is analogous to an n-way merge algorithm, though in this case a merge is not being performed, but rather, a set of calculations is being done.

Next, all of the query words are compared, searched, or processed vis-a-vis the first document 21, and a complete document "score" is computed for the first document 23. In performing a similarity search in accordance with the invention, it is desirable to keep a list of all the documents in the collection, or at least a list of all the documents that have been seen in the lists being processed. This is desirable in order to track the partial score of the documents. This list can be accessed at points corresponding to the document id portions of the document id_x - word frequency pairs being processed. Thus, as the list for each query word is processed the document list may be accessed at increasing (or decreasing) points, depending upon the ordering of the document ids.

The process is continued by retrieving the next document id - word frequency pairs for the next document 25 into the computer memory, and again processing all of the query words vis-a-vis the next document 27, and a new "score" is computed for the next document 28. The process is continued 30 until all of the documents have been processed. Once all the query words have been processed the fully computed or cumulative "scores" are sorted into rank order and the list displayed 31. Alternatively, in order to produce a sorted list immediately at the end of the process, each time a partial score is computed, the changed document score can be repositioned in the ranking as necessary.

Again, in contrast to previous techniques in which, if there was not sufficient memory in the system to keep the whole list in memory together with the portion of the query word's list of document id - word frequency pairs being processed, most of the document list was paged in from an external store, for comparison with each query word. In the technique of the invention, one query word's stream of pairs is switched to another to make all the calculations, for example, for the lowest document id in all of the various lists before going on to the second lowest document id and so forth. In accordance with the invention, there need only be enough memory to contain the entry for one document in the document list at a time, and for each query word, one entry in the list of pairs of document id - word frequency. Since for large document collections the list of documents will be very large this enables computation with a much smaller memory requirement than previous techniques.

It will be noted that it may be necessary to perform additional computation than previous techniques in making comparisons between the identifications of the current element in the various query word lists. However, this computation is inexpensive compared to disk input/output costs.

An alternative embodiment of the method of the invention is to process some number of documents (more than one) at the same time. This number of documents could be determined at run time based on the available memory, or at compile time based on the expected target machine. Each list of document id - word frequency pairs would be processed until the document identifications exceeded the current range of document identifications being processed. Then computation would move on to the next query word list. This variation decreases the amount of extra computation done, although it does not eliminate it entirely, and requires more memory, though not as much as previous approaches.

Claims

1. A method for performing a search of a plurality of documents for similarity to a query, comprising:
   (a) retrieving a first document;
   (b) determining a number of occurrences of said query in said first document;
   (c) retrieving a next document;
   (d) determining a number of occurrences of said query in said next document;
   (e) repeating steps (c) and (d) until each of said plurality of documents have been retrieved, and the number of occurrences of said query has been determined in each of said plurality of documents.

2. A method for performing a search of a plurality of documents for similarity to a plurality of query words, comprising:
   (a) retrieving a first document;
   (b) determining a number of occurrences of each of said plurality of query words in said first document;
   (c) retrieving a next document;
   (d) determining a number of occurrences of each of said query words in said next document;
   (e) repeating steps (c) and (d) until each of said plurality of documents have been retrieved, and the number of occurrences of each of said plurality of query words has been determined in each of said plurality of documents.

3. The method of claim 1 or 2 wherein said query comprises a number of query words.

4. A method for performing a search of a plurality of documents for similarity to a plurality of query words, comprising:
   (a) retrieving each of said documents in turn;
   (b) determining a number of occurrences of each of said plurality of query words in each of said documents in turn when each of said documents is retrieved.

5. A method for performing a search of a plurality of documents for similarity to a plurality of query words, comprising:
   (a) retrieving a first portion of said plurality of documents;
   (b) determining a number of occurrences of each of said plurality of query words in each document in said first portion of said plurality of documents;
   (c) retrieving a second portion of said plurality of documents;
   (d) determining a number of occurrences of each of said plurality of query words in each document in said second portion of said plurality of documents.

6. A method for performing a search of a plurality of documents for similarity to a plurality of query words, comprising:
   generating an index of entries for all words of all of said documents, each of said documents being identified by a document identifier, each entry containing a document identifier and a number of occurrences that a word appears in the identified document;
   for each document identifier, in turn, comparing each of said plurality of query words to each word coupled with of each document iden-

tifier.

7. The method of any of claims 3 to 6 further comprising ranking the documents according to the number of occurrences of said query words determined in each document.

8. The method of claim 7 further comprising producing a list of documents according to the document ranking.

9. The method of any of the preceding claims wherein said steps of retrieving a document comprises retrieving an image of the retrieved document into an electronic memory.

10. The method of any of the preceding claims wherein said step of determining a number of occurrences of each of said query words comprises generating a list of words contained within the retrieved document, and comparing the query words against the generated list of words.

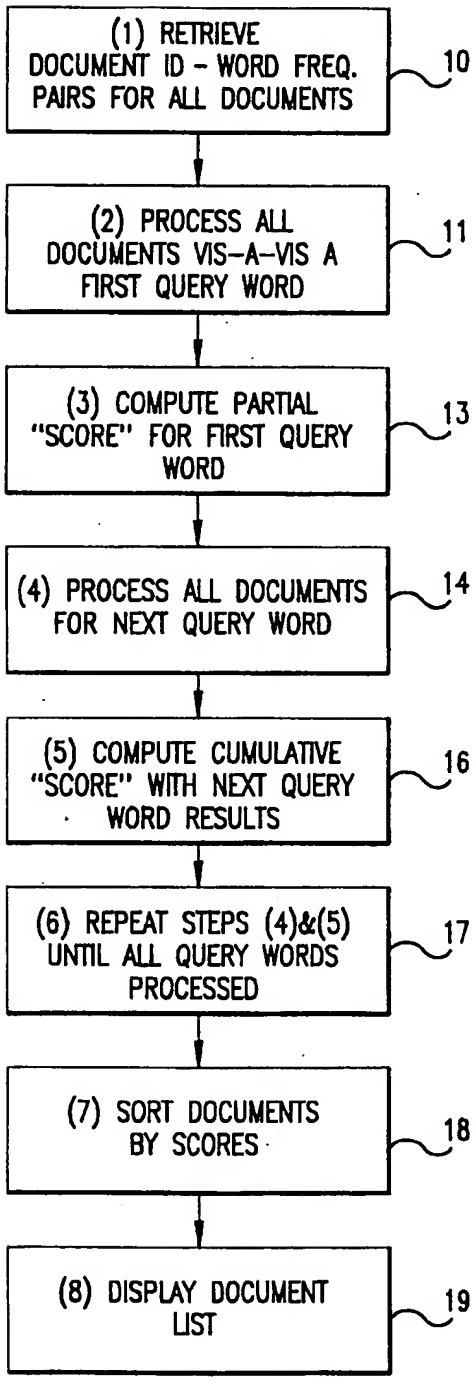

FIG.1
PRIOR ART

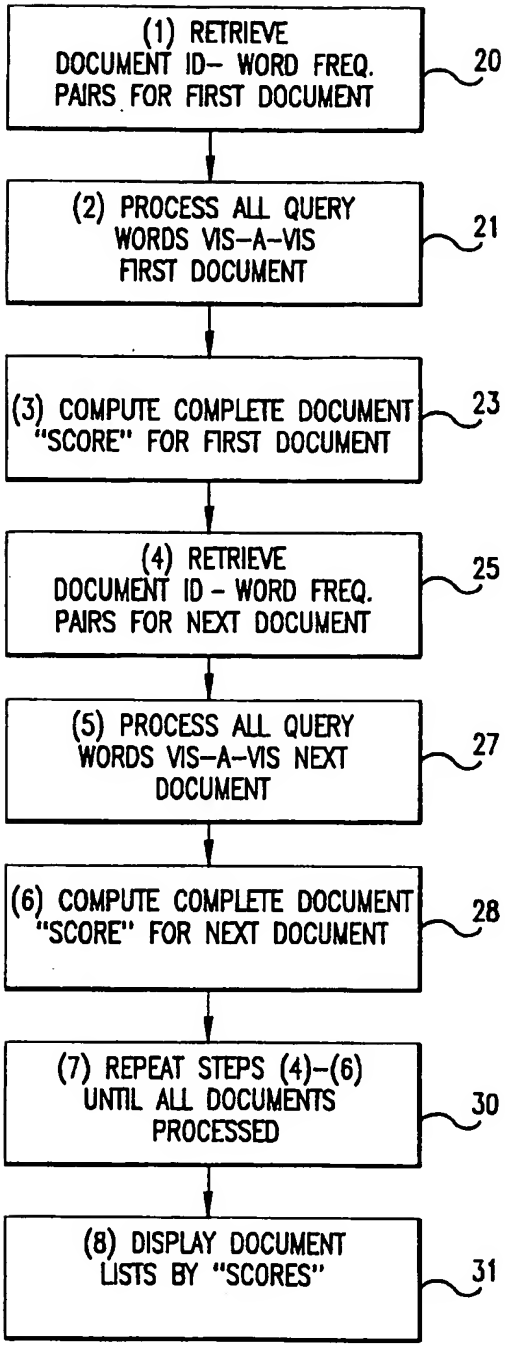
(1) RETRIEVE DOCUMENT ID– WORD FREQ. PAIRS FOR FIRST DOCUMENT
20
(2) PROCESS ALL QUERY WORDS VIS–A–VIS FIRST DOCUMENT
21
(3) COMPUTE COMPLETE DOCUMENT "SCORE" FOR FIRST DOCUMENT
23
(4) RETRIEVE DOCUMENT ID – WORD FREQ. PAIRS FOR NEXT DOCUMENT
25
(5) PROCESS ALL QUERY WORDS VIS–A–VIS NEXT DOCUMENT
27
(6) COMPUTE COMPLETE DOCUMENT "SCORE" FOR NEXT DOCUMENT
28
(7) REPEAT STEPS (4)–(6) UNTIL ALL DOCUMENTS PROCESSED
30
(8) DISPLAY DOCUMENT LISTS BY "SCORES"
31

FIG.2

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 93 30 7488

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.5)
Y	8TH ANNUAL INTERNATIONAL CONFERENCE ON COMPUTERS AND COMMUNICATIONS 22 March 1989 , USA pages 567 - 571 D. LUCARELLA : 'Heuristics to locate the best document set in information retrieval systems' * page 568, column 2, line 32 - line 44 *	1-10	G06F15/403
Y	PATENT ABSTRACTS OF JAPAN vol. 14, no. 238 (P-1050)21 May 1990 & JP-A-20 059 861 (NEC CORP.) 18 February 1990 * abstract *	1-10	
X	EP-A-0 501 416 (HITACHI ,LTD) * page 4, line 35 - page 5, line 7; figures 3,5 *	1-10	
A	COMPUTER JOURNAL vol. 35, no. 3 , June 1992 , LONDON GB pages 279 - 290 H. TURTLE & B. CROFT : 'A Comparison of Text Retrieval Models' * page 284, column 2, line 1 - page 285, column 1, line 32 *	1-10	TECHNICAL FIELDS SEARCHED (Int.Cl.5) G06F
A	EP-A-0 304 191 (IBM CORP.) * abstract; claims 1-7,9 *	1-10	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	7 January 1994	Fournier, C

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)